The Systems Toxicology Computational Challenge:

Marker of Exposure Response Identification

Overview

Carine Poussin, PhD
July 11th 2016, ISMB, Orlando
sbv IMPROVER in a Nutshell
IMPROVER: Industrial Methodology for Process Verification in Research

- Project initiated 6 years ago and funded by Philip Morris International R&D
- Aims to verify methods & data in systems biology / toxicology
- Verifies individual methods using double blind performance assessment

Verification of systems biology research in the age of collaborative competition

- Pablo Meyer, Leonidas G Alexopoulos, Thomas Bork, Andrea Califano, Carolyn R Ciof
- Alberto de la Fuente, David de Graaf, Alexander J Hartemink, Julia Hoeng, Nikolai V Ivanov
- Heinz Koeppf, Rume Linding, Daniel Marbach, Raquel Norel, Manuel C. Peteji, J Jeremy Rice
- Ajay Royyuru, Frank Schacherer, Jorg Sprenger, Katrin Stolze, Dennis Vitkup & Gustavo Stolovitzky

Collaborative competitions in which communities of researchers compete to solve challenges may facilitate more rigorous scrutiny of scientific results.

Nature Biotechnology 2011 Sep 8;29(9):811-5
Previous sbv IMPROVER crowd-sourcing challenges

1. Diagnostic Signature Challenge (2012)
The identification of gene expression signatures and computational methods for diagnostic classification

2. Species Translation Challenge (2013)
The translatability of biological system perturbations across species

The verification and enhancement of biological causal networks representative of various biological processes

Challenge outcome including best performing methods and lessons learned were:
• Presented in symposium
• Published in peer-reviewed journals

https://sbvimprover.com/sbv-improver/publications/
2016 - The Systems Toxicology Computational Challenge

https://sbvimprover.com/challenge-4
Overview

To develop a classification approach that identifies a blood gene signature capable of associating subjects to the correct exposure group
The Systems Toxicology Computational Challenge
Marker of Exposure Response Identification

Goal: to develop **inductive** blood gene signature-based classification models to predict **smoking exposure** or **cessation status**

**Sub-challenge 1 (SC1)**
Human blood signature as exposure response marker

**Sub-challenge 2 (SC2)**
Species translatable blood gene signature as exposure response marker

**Human** signatures

**Species-independent** signatures

**Robust** and **sparse** gene signatures → do not exceed 40 genes
Classification models

**Inductive classification model** as opposed to transductive

Apply a trained classification model to predict class label of any new individual sample (no retraining or use training data)
Training and Test data

**TRAIN**

- GEX DATA + CLASS LABELS

  Use of additional public and private datasets allowed

**TEST**

1. Randomized samples
2. Split test data in 2 subsets sequentially released

- Subset 1
  - Data released 20 Nov 2015
  - Submission closed 7 Apr 2016
- Subset 2
  - Data released 11 Apr 2016
  - Submission closed 29 Apr 2016

- Common set of 64 samples

**GOLD STANDARD**

= CLASS LABEL

**TRAIN INDUCTIVE MODEL**

**PREDICT CLASS LABEL**

- APPLY
Study datasets

Gene expression data generated from human and mouse blood samples

**TRAIN**

- **dset1**
  - **S** (109)
  - **FS** (57)
  - **NS** (58)

- **dset4**
  - **3R4F** (40)
  - **Cess** (27)
  - **Sham** (45)

**TEST**

- **dset2**
  - **S** (27)
  - **FS** (26)
  - **NS** (28)

- **dset5**
  - **3R4F** (12)
  - **Cess** (8)
  - **Sham** (13)

**SC1** ✓✓ ✓✓

**SC2** ✓✓ ✓✓

**Task 1:**

**signature-based classification model 1**

*(S vs NCS)*

**Task 2:**

**signature-based classification model 2**

*(FS vs NS)*

S/3R4F: Smokers / 3R4F (exposure to smoke from a reference cigarette)

FS/Cess: Former smokers / Cessation

NS/Sham: Never smokers / Sham

NCS: Non-current smoker

Freedom to use two separate models for 2-class prediction for each step, or directly a 3-class prediction model

Submissions and Timelines

Submissions for each sub-challenge
- 4 class label prediction files
- 2 gene lists
- 1 write up

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Smoker</th>
<th>Non-current smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>P1</td>
<td>P2</td>
</tr>
<tr>
<td>Sample 2</td>
<td>0.95</td>
<td>0.05</td>
</tr>
<tr>
<td>Sample 3</td>
<td>0.94</td>
<td>0.04</td>
</tr>
<tr>
<td>....</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample M</td>
<td>0.85</td>
<td>0.15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Former Smoker</th>
<th>Non-current smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>P1</td>
<td>P2</td>
</tr>
<tr>
<td>Sample 2</td>
<td>0.95</td>
<td>0.05</td>
</tr>
<tr>
<td>Sample 3</td>
<td>0.94</td>
<td>0.04</td>
</tr>
<tr>
<td>....</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample M</td>
<td>0.85</td>
<td>0.15</td>
</tr>
</tbody>
</table>

- \( P_x \) = confidence value that a sample belongs to a class
- \( P1+P2=1 \); \( P1 \neq P2 \)